Citation: Clin Transl Sci (2018) 00, 1–6; doi:10.1111/cts.12558

ARTICLE

How Often Do Safety Signals Occur by Chance in First-in-Human Trials?

Gemma L. Clayton[1,*], Asher D. Schachter[3], Baldur Magnusson[2], Yue Li[2] and Laurence Colin[3]

Clinicians working on first-in-human clinical studies need to be able to judge whether safety signals observed on an investigational drug were more likely to have occurred by chance or to have been caused by the drug. We retrospectively reviewed 84 Novartis studies including 1,234 healthy volunteers receiving placebo to determine the expected incidence of changes in commonly measured laboratory parameters and vital signs, in the absence of any active agent. We calculated the frequency of random incidence of safety signals, focusing on the liver, cardiovascular system, kidney, and pancreas. Using the liver enzyme alanine aminotransferase (ALT) as an example, we illustrate how a predictive model can be used to determine the probability of a given subject to experience an elevation of ALT above the upper limit of the normal range under placebo, conditional on the characteristics of this subject and the study.
Clin Transl Sci (2018) **00**, 1–6; doi:10.1111/cts.12558; published online on yyyy-mm-dd.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
✔ Little information is available in the literature on the expected variations in laboratory values and vital signs in a clinical study setting, when subjects receive placebo. A few reviews have mentioned unexpectedly high rates of ALT elevations in subjects taking placebo, but were based on small data sets and did not adjust for individual characteristics.

WHAT QUESTION DID THIS STUDY ADDRESS?
✔ Our goal was to provide a tool for clinicians working on first-in-human studies, to quantify whether safety signals observed were likely the result of chance or the compound under investigation.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
✔ We provided reference incidence rates under placebo for commonly measured safety parameters. We built a predictive model for ALT elevations that can be used to quantify precisely how likely an event is due to chance, conditionally on the characteristics of the subject and the study.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?
✔ This work should help teams identify safety signals earlier and with greater accuracy.

A report published in 2014 revealed that about half of all US Food and Drug Administration (FDA) rejections and delayed approvals in recent years were due, at least in part, to safety deficiencies,[1] with cardiovascular and hepatic issues being the most common concerns. Quantitative tools that identify and characterize safety issues earlier in the life cycle of investigational compounds would have a large impact on the efficiency of drug development, since the most costly phases of drug development are phase II and phase III.[2] Yet early phase studies often lack quantitative evaluations of safety data.

First-in-human (FIH) studies offer the first opportunity to test ascending doses of a new treatment in healthy volunteers. The primary objective of these studies is to investigate the safety and tolerability of a new drug. Due to the typically small sample sizes of cohorts, safety signals are often difficult to interpret, particularly in the absence of a robust placebo group. Previous attempts at quantifying the background rate of liver enzyme elevations in placebo-treated healthy individuals[3,4] were limited by small data sets and did not take into account the demographic and background characteristics of the healthy volunteers in their estimation of the incidence rates. For most other safety events, no reference rates are available at all.

METHODS

Description of the data

We retrospectively reviewed studies in the Novartis database that involved healthy volunteers and were completed before 2016. All clinical studies included in this article were sponsored by Novartis and reviewed by an Institutional Review Board (IRB). Of those, we excluded 67 studies that did not involve placebo and 44 studies that used a crossover design. There were 11 studies for which the laboratory and vital signs data were not readily available. All the placebo data from the remaining 77 studies were pooled. Among these 77 studies, 69 were single-escalation or multiple-escalation designs, of

[1]Bristol Medical School, Population Health Sciences, University of Bristol, U.K; [2]Novartis Pharma AG, Basel, Switzerland; [3]Novartis Institute for Biomedical Research, Cambridge, Massachusetts, USA. *Correspondence: Gemma L Clayton (gemma.clayton@bristol.ac.uk)
Received 29 January 2018; accepted 30 March 2018; published online on yyyy-mm-dd. doi:10.1111/cts.12558

Table 1 Demographics of the placebo database

***N* = 1,234 subjects**	***n***	**%**
Sex: Male (*n*, %)	1,017	82.4
Ethnicity (*n*, %):		
White	662	53.6
Hispanic or Latino	259	21.0
Asian	141	11.4
Black or African American	125	10.1
Other	47	3.8
Continent (*n*, %):		
America/Canada	707	57.3
Europe	399	32.3
Australia	40	3.2
Asia	88	7.1
Age (years) (Median, Q1-Q3)[a]	32	25–43
Height (cm) (Median, Q1-Q3), $N = 1{,}212$	175.0	168.2–181.0
Weight at baseline (kg) (Median, Q1-Q3)	77.1	68.0–85.8

[a]Q1 = first quartile (25th percentile), Q3 = third quartile (75th percentile). Age ranged from a minimum of 18 to a maximum of 78 years. Height ranged from a minimum of 143.8 to a maximum of 199.0 cm. Weight ranged from a minimum of 47.7 to a maximum of 116.1 kg.

which 10 were conducted in Japanese subjects and one in a Chinese subject. The number of placebo subjects per study varied between 3 and 57, with a mean of 16.03 per study. The number of postbaseline observations per study varied between one and 18, with a mean of 5.65.

We selected the following routinely measured safety parameters for our review:

- Liver safety: alanine aminotransferase (ALT), aspartate aminotransferase (AST), bilirubin.
- Cardiovascular safety: the Fridericia-corrected QT interval (QTcF), standing systolic blood pressure (SBP), heart rate (HR).
- Renal safety: serum creatinine.
- Pancreatic safety: lipase, amylase.

While normal laboratory ranges are known for all of the parameters listed above, incidences of randomly occurring values outside of the normal ranges for healthy subjects receiving placebo in the setting of a clinical study are not known.

The pooled database included 1,234 subjects with available measurements in at least one of the categories above. The demographic characteristics of the population are presented in **Table 1**.

We present the raw incidences of various safety events in the pooled database and explain how these can be used to give a preliminary assessment of whether signals observed during the use of an investigational drug are in line with the expected incidence with placebo.

A model that adjusts for potential study differences and subject characteristics provides a more precise assessment. We use the liver enzyme ALT (U/L) as an example to illustrate the potential application of this method.

Demographic and background information included a subject's age, gender, ethnicity, continent, height, weight, and baseline value. Baseline ALT was taken postrandomization and prior to the placebo being given. Due to differences between assays used, the upper limit of the normal range (ULN) varied slightly and we therefore normalized the ALT values by the corresponding ULN value and used ALT/ULN as the response variable for modeling purposes. Histograms of continuous covariates were produced to check for normality and potential outliers. Log transformation was applied to baseline ALT and age. Baseline ALT was calculated in the unit of ULN and then log transformed to obtain an approximately normally distributed variable. Age was also log transformed for the same reason. Log transformation was initially attempted for the number of samples; however, this did not improve the model fit and therefore the original scale was used to aid interpretation. Weight was approximately normally distributed and therefore no transformation was considered necessary. All variables were standardized to a scale with a mean of 0 and standard deviation of 1. The standardization was done via the typical approach, i.e., subtracting the mean of the variable from each of the individual (subject) values and dividing by the overall standard deviation.

Description of the model

We modeled the probability of a subject developing an event of ALT $>$ ULN at least once during a study. To this end, we fitted a multilevel logistic regression model as follows.[5] For subject i in study j we define the event $y_i = 1$ if the subject had at least one ALT measurement exceeding the ULN during the study. The probability of this event was modeled as Bernoulli[logit$^{-1}(\alpha_j + X_i\beta)$], where α_j represents the study-specific intercept used to account for between-trial variation, X_i is a vector of covariates specific to subject i (including the number of postbaseline samples for subject i), and β is a vector of covariate parameters. The model was fitted in the Bayesian framework using the following weakly informative priors: $\alpha_j = \alpha + u_j$, with $\alpha \sim$ Cauchy(0, 10), $u_j \sim N(0, \tau^2)$, $\tau \sim$ Exponential(1), and $\beta \sim$ Cauchy(0, 2.5). Since a logistic regression model was fit, model coefficient estimates are reported as the log transformed odds ratios of the posterior means.

The baseline covariates described in **Table 1** were evaluated for inclusion in the model, using a forward selection that calculates the difference in deviance for nested models. The model with the smallest expected log pointwise predictive density was selected.[6] Model fit was checked by comparing posterior predictive distributions to observed values. Model fitting was done using Stan[7] via the R library (v. 3.4.1) RStanArm (v. 2.15.3).[8] No time-trend was observed in the longitudinal data and hence a time slope was not included in the above model. Instead, the longitudinal aspect of the data was taken into account by including the number of postbaseline assessments as a covariate in the model.

The "virtual placebo twin"

From this model, we can derive the subject-specific probability of experiencing an event, conditionally on this subject's covariates. For subject i in our data set, we would condition on the study-level effect α_j, while for a new subject this probability would be obtained by integrating over the study effect distribution[9,10] (see **Supplementary Material** for more

detail). We can also derive a marginal (population average) probability of a subject to experience an event by averaging over the distribution of covariates. Due to the nonlinearity of the model, the conditional probability of a subject with mean covariate values is not expected to match the population mean marginal probability. In other terms, taking the mean before or after the logit transformation will not yield the same results, as it is always the case for generalized linear models with a nonlinear link function. In this setting, because we are mostly interested in predictions for specific subjects in a new study (for whom we know the baseline covariates), the conditional probability is of most interest. For a subject receiving an active drug in a new study experiencing an ALT $>$ ULN event, we will produce the predicted probability of this subject also experiencing this event had he/she received placebo instead of active drug, based on the model described above. We will call this prediction the probability of a "virtual placebo twin" to experience the event. Whether or not the virtual placebo twin is also likely to have experienced an event will determine whether the investigational drug is likely to have caused the event or not.

Finally, we can combine the individual subjects' probabilities to estimate the probability that at least one subject experiences one event in a cohort of size *n*. This can be done by using conditional probabilities (a different one for each subject) if subject-specific covariates are available, or marginal probabilities (identical for each subject) if subject covariates are not available.

RESULTS

The raw incidence of various safety signals in our pooled database of healthy volunteers receiving placebo are shown in **Table 2**, by target organ. This table can be used to judge how frequently random safety findings occur in a healthy population. For example, the table's data indicate that increases in heart rate (HR) by more than 20 beats per minute from baseline occur in about 14.16% of healthy subjects receiving placebo. Therefore, in a cohort of six subjects, observing HR increases of this magnitude in two subjects receiving active drug wouldn't necessarily be a concern, since it is not unlikely to happen in the same population receiving placebo: the probability of observing at least two subjects with an event in a cohort of six, if the event truly occurs with a 14.16% probability, is 20% (from the binomial distribution, see **Table 3**). On the other hand, elevations of amylase above 2 times the upper limit of normal only occur in 0.33% of healthy subjects receiving placebo. In this case, observing just one subject in a cohort of six subjects with elevations of this magnitude raises concerns about the investigational treatment, since this is unlikely to happen (2% chance) under placebo (**Table 3**).

While these raw incidence rates are helpful in providing a quick assessment of the likelihood for a safety signal to be caused by the active compound under investigation, more accurate answers can be given with a model adjusting for differences between study and individual subject characteristics.

Using the event ALT $>$ ULN as an example, the raw incidence of elevations in healthy volunteers taking placebo is 6.24%, as shown in **Table 2**. However, this prediction varies substantially across individuals. Following the model selection procedure described earlier, a logistic model of the probability of a subject to experience at least one ALT $>$ ULN event during the study was built, and the final covariates and model coefficients are shown in **Table 4**.

The between-study variance is 1.20, which represents 27% of the residual variance. The impact of age and weight on the probability of a subject developing an ALT elevation above ULN during a study is small in magnitude, and perhaps counterintuitively, is negative: increasing age and weight is associated with reduced risk of a random ALT elevation under placebo. Of note, including sex did not improve the model significantly; however, this could be due in part to the fact that 82% of the subjects in our pooled database were male. The most important predictor of an event is ALT at baseline, but the number of samples collected during the study also influences the likelihood of observing an event. Therefore, as highlighted in **Figure 1a**, subjects with baseline ALT close to ULN and in studies with $>$8 samples taken have a much higher probability of an ALT $>$ ULN.

Figure 1a shows the distribution of individual predicted probabilities of ALT $>$ ULN for all subjects in the data set. For a typical subject with average covariate values (baseline ALT value of 21.5 U/L, ULN of 55 U/L, age 32.8 years with a weight of 77.1 kg, and 5.65 postbaseline observations over the study), the probability of developing an ALT $>$ ULN is 2.4%. This is substantially lower than the population mean of 6.2%, because the distribution of probabilities is skewed to the right, and higher-risk subjects drive the average up. The most influential covariate is a subject's baseline ALT value and how close it is to the ULN. **Figure 1b** shows that the predicted probability of an ALT $>$ ULN event varies from 2% for a baseline of 20 U/L to 19% for a baseline of 40 U/L, controlling for all other covariates.

To illustrate how these findings can be applied in practice in the setting of a dose escalation study of a new investigational drug, consider a hypothetical cohort of six active-treated subjects, where we randomly chose some high ALT values from the cohort at baseline, in order to illustrate the substantial impact of this variable. Other baseline characteristics were selected at random within the interquartile range. Applying the model described above generates the probability that each subject's placebo twin would have had an event (i.e., the probability that this subject would have developed an event had he/she been receiving placebo), as shown in the last column of **Table 5**.

Based on these baseline characteristics, the probability of observing at least one ALT $>$ ULN by chance in this cohort of six subjects is 57%. Therefore, observing one event in this specific cohort wouldn't raise particular concerns about the liver safety of the active drug:

$$
\begin{aligned}
&P(\text{at least one event}) = 1 - P(\text{zero event}) \\
&= 1 - (0.986 * 0.971 * 0.938 * 0.843 * 0.866 * 0.661) \\
&= 0.57
\end{aligned}
$$

Table 2 Raw incidence (unadjusted for study effect) of safety signals in pooled database of placebo-treated healthy volunteers

Target organ	Safety event	Raw incidence (number of subjects with at least one event /total number of subjects) in pooled early safety studies	Estimated incidence rate (%)
Liver	ALT > ULN	77/1,234	6.24
	ALT > 2 x ULN	10/1,234	0.81
	ALT > 3 x ULN	4/1,234	0.32
	Bilirubin > ULN	92/1,180	7.80
	Bilirubin > 2 x ULN	36/1,180	3.05
	Bilirubin > 3 x ULN	30/1,180	2.54
	ALT or AST > 3 x ULN; & Bilirubin > ULN	0/1,234	0
Cardiovascular system[a]	QTcF change > 60 ms & QTcF < 500 ms	7/1,028	0.68
	QTcF change > 60 ms & QTcF ≥ 500 ms	0/1,028	0
	HR increase > 20 bpm	165/1,165	14.16
	Standing SBP increase > 20 mmHg	64/790	8.10
Kidney[b]	Serum creatinine increase > 50%	0/1,234	0
Pancreas	Lipase > 1.5 x ULN	34/1,125	3.02
	Lipase > 3 x ULN	7/1,125	0.62
	Amylase > 2 x ULN	4/1,195	0.33

ALT, alanine aminotransferase; AST, aspartate aminotransferase; QTcF, Fridericia-corrected QT interval; HR, heart rate; standing SBP, standing systolic blood pressure (SBP) which is when blood pressure is taken when the subject is standing up. Units: ULN, upper limit of normal; bpm, beats per minute; mmHg, millimeter of mercury.

[a]Baseline QTcF ranged from 347 to 481 ms, with a mean of 398.0 (SD = 12.5). Baseline HR ranged from 37 to 125 bpm, with a mean of 62.3 (SD = 10.7). Standing SBP ranged from 86 to 168 mmHg, with a mean of 119.5 (SD = 11.9).

[b]Baseline serum creatinine ranged from 35 to 168 umol/L, with a mean of 81.2 (SD = 14.4). The ULN varied across studies with a median of 112 umol/L (IQR 106 to 115).

Table 3 Hypothetical situations in first-in-human studies and the corresponding probability to observe the same events under placebo

Safety event	Number of subjects under active drug with an event	Rate of event occurrence under placebo	Probability to observe 2/6 events under placebo	Conclusion
HR > 20 bpm from baseline	2/6	14.16%	20%	Situation is *not* unlikely to have happened under placebo
Amylase > 2 × ULN	1/6	0.33%	2%	Situation unlikely to have happened under placebo

The probability of at least two placebo twins experiencing an ALT > ULN event is 15%. Therefore, observing two such events on active drug in this cohort may raise suspicions that would require further confirmation. **Figure 2** illustrates the probability of observing at least 1, 2, 3, 4, 5, or 6 placebo twins with an event in this cohort.

DISCUSSION

This is the first known large review of the expected frequency of random safety findings in placebo-treated healthy volunteers. The reference event rates provided for parameters relating to the safety of the liver, cardiovascular system, kidney, and pancreas will provide valuable insight for clinical teams assessing the safety of investigational compounds in early phase studies. Using the liver enzyme ALT as an example, we showed how predictive models can provide a more precise assessment of the chance occurrence of a safety signal in a subject. We developed a logistic model for the probability of a subject developing an ALT > ULN event during a study, while taking placebo. This model showed that the most important factor influencing the chance of a random event is the ALT value at baseline. In our data set, the mean probability estimate of a healthy volunteer to develop an ALT > ULN event under placebo is 6.2%. For a typical subject with average baseline characteristics (baseline ALT value of 21.5 U/L), this probability is 2.4%. If the baseline ALT doubles to 40 U/L, this probability increases to 19%. This illustrates that caution should be taken when interpreting ALT elevations in cases where the baseline value is higher

Table 4 Final model coefficient estimates for the logistic model of the probability of a subject experiencing an event of ALT > ULN

Model coefficients[a]	Posterior median (log odds)	95% credible interval
Intercept	–4.16	(–4.88, –3.59)
Baseline ALT (U/L) (log transformed)	1.65	(1.30, 2.05)
Number of postbaseline samples taken	0.53	(0.22, 0.86)
Age in years (log transformed)	–0.33	(–0.64, –0.03)
Weight in kg	–0.31	(–0.65, –0.003)
Between-study variability (τ^2)	1.20	(0.39, 2.84)

[a]All variables are standardized to a scale with a mean of 0 and a standard deviation of 1.

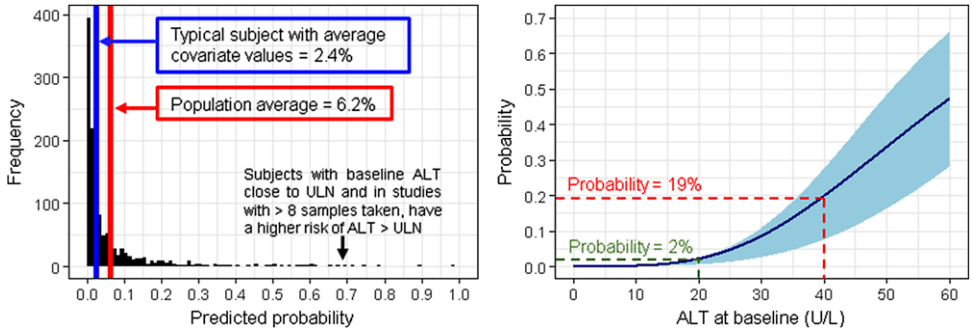

Figure 1 Model based predicted probabilities of a subject to experience an event of ALT > ULN. **(a)** Distribution of model-predicted individual probabilities of ALT > ULN for all subjects in the data set. **(b)** Predicted probability (and interquartile range) of an ALT > ULN for a subject depending on baseline ALT (other covariates fixed at the mean population values). For more details on how these probabilities were calculated, see **Supplementary Material**.

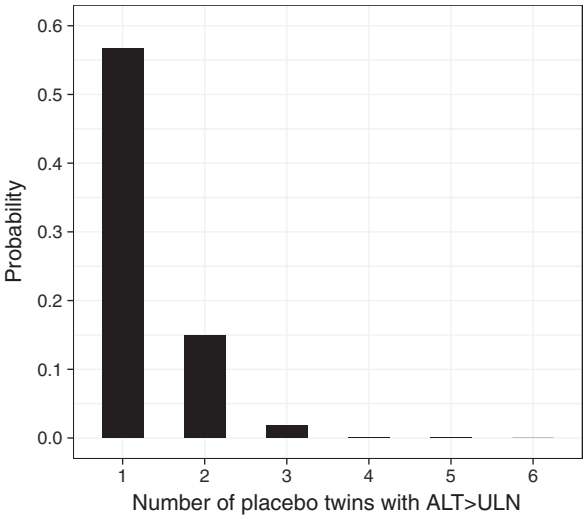

Figure 2 Probability of observing a given number of subjects with an event by chance in the hypothetical cohort described by **Table 5**.

than usual. Sponsors may elect to recruit only subjects with a predicted probability of a random elevation that is lower than 10% in FIH studies, since for those subjects it will be easier to attribute any emerging liver safety signal to the investigational drug. Previous attempts at quantifying the expected incidence of ALT elevations in healthy volunteers taking placebo[3] have reported a higher ALT > ULN event rate of 20.4%. This is probably due to the use of a small data set and the inclusion of less healthy subjects with higher baseline ALT levels. This highlights further the importance of controlling for baseline when making these predictions.

There are a few limitations to this work. The first one is the choice of modeling the ALT > ULN on a binary scale. We first tried to model ALT on a continuous scale, as this should theoretically make more efficient use of the data than modeling a binary event such as ALT > ULN directly. However, we could not find a distribution that described the continuous data adequately (especially the tails, which are crucial for this exercise) and a model that provided unbiased predictions of the number of ALT > ULN events. Another limitation is that some of the events described in the text as "random elevations" may be partially explained by factors that were not captured in the database. Differences in domiciling and subject management between studies may explain some of these differences, and unmeasured medical history or other study-specific design features (such as food intake, etc.) may explain others. For example, while our database did not capture this information, we know that most subjects in the data set were domiciled at least for the first 3–5 days of the study and had normal access to food three times a day. Data collection in FIH studies is likely not optimal for this type of exercise. Each company may have different standards for FIH protocols and the numbers observed in studies sponsored by companies other than Novartis may look slightly different for this reason.

Before using this database, we recommend checking that the range of subjects' baseline characteristics in the new study is appropriately represented in the historical data. For

Table 5 Hypothetical cohort of active-treated subjects and individual model-based predictions for each individual's placebo twin

Subject	Baseline ALT (U/L)	ULN (U/L)	Number of postbaseline samples taken	Age (years)	Weight (kg)	Predicted probability ALT>ULN for placebo twin
1	16	55	6	22	75	1.4%
2	21	55	6	32	78	2.9%
3	28	55	6	47	70	6.2%
4	35	55	6	25	80	15.7%
5	40	55	6	52	76	14.4%
6	50	55	6	35	77	33.9%

example, if the average age in the new study is above the 99% percentile of age in the pooled database, the model predictions may not be reliable, since they will be heavily dependent on 12 subjects or less, which is <1% of the cohort.

This pooled data set should be updated annually to reflect the most up-to-date information and these results may therefore change with time.

We showed that a predictive model can be used to create "virtual placebo twins," i.e., subjects with the same baseline characteristics, but who would have received placebo: for every subject experiencing an event under an investigational drug, the model will predict the likelihood of his/her "virtual placebo twin" experiencing the same event. The lower this model-predicted probability, the higher the chance that the drug under investigation is causing the issue. In early studies, decisions are made in the context of all available data (including preclinical evidence, pharmacokinetic data, etc.) and this tool is not intended to lead teams to ignore this complexity of information; rather, it should be viewed as a way to consider one piece of the complex array of data (namely, the rates of laboratory abnormalities) in a more objective and quantitative way.

Finally, this work could be extended to patient populations (we think it would have particular value for rare populations, for example, pediatric studies). This would come with additional complications, since patient studies often do not share similar designs and inclusion criteria, unlike FIH studies of healthy volunteers. Nevertheless, we think the outcome of this exercise could help distinguish effects related to the drug under study from the underlying disease, in populations where the effect of placebo has been poorly studied. In addition, this work could also be extended by using historical placebo data from patient studies in the same or similar disease area to inform a new patient study.

By using the large amounts of placebo data collected in healthy volunteers over decades of clinical investigations, companies can contribute to increasing the quality of safety decision-making in early phase clinical trials. We can quantify with higher precision the expected frequency of random safety signals in FIH studies and separate real signals from the noise.

Funding. Funding was provided by Novartis Pharma AG.

Conflict of Interest. L.C., A.S., B.M., and Y.L. are employees and shareholders of Novartis Pharma AG.

Author Contributions. L.C., G.C., and B.M. wrote the article; A.S. and L.C. designed the research; L.C., G.C., B.M., and Y.L. performed the research; G.C. analyzed the data.

1. Sacks, L.V., *et al.* Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000–2012. *JAMA* **311**, 378–384 (2014).
2. Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. How much do clinical trials cost? *Nat. Rev. Drug Disc.* **16**, 381–382 (2017).
3. Rosenzweig, P., Miget, N. & Brohier, S. Transaminase elevation on placebo during Phase I trials: prevalence and significance. *Br. J. Clin. Pharmacol.* **48**, 19–23 (1999).
4. Merz, M., Seiberling, M., Hoxter, G., Holting, M.L. & Wortha, H.P. Elevation of liver enzymes in multiple dose trials during placebo treatment: are they predictable. *J. Clin. Pharmacol.* **37**, 791–798 (1997).
5. Molenberghs, G. & Verbeke, G. *Models for discrete longitudinal data.* New York: Springer; 2005.
6. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comp.* **27**, 1413–1432 (2017).
7. Carpenter, B. *et al.* Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–29 (2017).
8. Stan Development Team. RStanArm: Bayesian applied regression modeling via Stan. R package version 2.15.3., (2017).
9. Skrondal, A. & Rabe-Hesketh, S. Prediction in multilevel generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **172**, 659–687 (2009).
10. Pavlou, M., Ambler, G., Seaman, S. & Omar, R.Z. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *Bmc Med. Res. Methodol.* **15**, XX (2015).

Supplementary information accompanies this paper on the *Clinical and Translational Science* website. (www.cts-journal.com)